This PDF is a selection from an out-of-print volume from the National Bureau of Economic Research

Volume Title: Social Experimentation

Volume Author/Editor: Jerry A. Hausman and David A. Wise, eds.

Volume Publisher: University of Chicago Press

Volume ISBN: 0-226-31940-7

Volume URL: http://www.nber.org/books/haus85-1

Publication Date: 1985

Chapter Title: Macroexperiments versus Microexperiments for Health Policy

Chapter Author: Jeffrey E. Harris

Chapter URL: http://www.nber.org/chapters/c8375

Chapter pages in book: (p. 145 - 186)

4 Macroexperiments versus Microexperiments for Health Policy

Jeffrey E. Harris

4.1 Introduction

In social *micro*experiments, the experimenter assigns treatments and gauges responses at the individual level. The response of each individual is assumed to be independent and small in comparison to the market or social system.

In social *macro*experiments, treatments are assigned at the group, community, or market level. The responses of entire social units, as well as of individuals within each unit, are the objects of interest. The responses of the individuals within each unit are correlated (Rivlin 1974; Mosteller and Mosteller 1979).

Economists and other social scientists have spent disproportionately too much effort on the design and interpretation of microexperiments. The potential value and limitations of macroexperiments have not been adequately characterized. Accordingly, we need to develop a new science of macroexperimental design and to articulate more carefully the trade-off between micro and macro designs as guides to public policy.

My argument is framed within the context of health-policy experiments. I concentrate on two policy issues: the effect of changes in health-insurance coverage on the demand for medical care and the effect of life-style intervention on the risk of coronary heart disease (CHD).

In section 4.2, I point out several problems in the design, implementa-

Jeffrey E. Harris is professor of economics, Massachusetts Institute of Technology, and research associate, National Bureau of Economic Research.

This paper was presented at the NBER Conference on Social Experimentation, Hilton Head Island, 5–7 March 1981. The work was supported by Public Health Service Research Grant no. DA–02620 and Research Career Development Award no. DA–00072. Valuable criticisms by Stephen Fortmann, Victor Fuchs, Samuel Greenhouse, Emmett Keeler, Joseph Newhouse, and Roger Sherwin are gratefully acknowledged.

tion, and interpretation of health-policy microexperiments. These include subject selection and attrition, anticipatory responses, Hawthorne effects, and ethical constraints on individual randomization. Although the results of microexperiments may elucidate certain mechanisms of individual behavior, they may not reveal the total, market equilibrium effects of policy alternatives.

Section 4.3 considers how macroexperiments may resolve these microexperimental difficulties. Because macroexperimentation can be less intrusive upon individuals, these experiments may avoid the potential selection and attrition biases, Hawthorne effects, and ethical constraints characteristic of microexperiments. Most important, macroexperiments can be more useful for evaluating the total market and social-system effects of policy options.

In section 4.4, I discuss two serious limitations of macroexperimentation. First, intervention at the market or community level reduces the statistical power of the experiment and, in some cases, threatens its external validity. Second, the macroexperimenter may encounter significant political and administrative obstacles to randomization.

Section 4.5 considers how these defects of macroexperimentation might be avoided. Decentralization of macroexperiments, along with experimental blocking, is suggested as a means of improving statistical power and overcoming administrative barriers to randomization. Time-series experiments, crossover designs, as well as mixtures of micro and macro designs, are considered. To resolve questions of external validity, I show how the results of different macroexperiments might be combined.

Throughout the analysis, I focus on the experience of two microexperiments—the Rand Health Insurance Study (Newhouse 1974) and the Multiple Risk Factor Intervention Trial (Multiple Risk Factor Intervention Trial Group 1976a, 1976b)—and one macroexperiment—the Stanford Heart Disease Prevention Program (Farquhar 1978). Several other macro-experiments in life-style intervention are in progress or under consideration.[1] But no bona fide macroexperiment in health insurance or in medical-care utilization has been undertaken. One goal of section 4.5 is to suggest how such experiments might be executed.

This paper is not a broad endorsement of macroexperimentation for health policy. It does not advocate the abandonment of microexperiments. Nor do I envisage a strict choice between micro and macro designs. But in many cases, precise microestimates of only one or two

1. These are the Stanford Five-City Project (Hulley and Fortmann 1980); the North Karelia Project (Puska et al. 1978); the Minnesota Heart Health Program; the Pawtucket Heart Health Program; the European Collaborative Heart Disease Prevention Project (WHO European Collaborative Group 1974; Rose et al. 1980); and the Pennsylvania County Health Improvement Program (Stolley 1980).

parameters of a problem do not justify our plunging into full-scale policies. Less precise macroassessments of the total impact of contemplated policies may then be warranted.

4.2 Problems with Microexperiments

First, I set forth the background of two microexperiments in health policy.

4.2.1 The Multiple Risk Factor Intervention Trial (MRFIT)

Epidemiologists have repeatedly shown that high blood pressure, elevated blood cholesterol, and cigarette smoking are independent, powerful predictors of an individual's risk of fatal and nonfatal events of coronary heart disease (Truett, Cornfield, and Kannel 1967). Men and women who spontaneously quit smoking incur a lower risk of subsequent coronary events than continuing smokers (Friedman et al. 1981). These findings have been derived from the natural histories of various study populations (for example, residents of Framingham, Massachusetts). To assess the causal nature of such predictive relationships, and to gauge the reversibility of the disease process, it would be logical to attempt to reverse each of the above risk factors in a randomized experiment.

Separate clinical trials have been instituted to lower blood cholesterol, to treat hypertension, and to induce smoking cessation (Davis and Havlik 1977; Hypertension Detection and Follow-up Program Cooperative Group 1979a, 1979b; Rose and Hamilton 1978). The difficulty with such single-factor experiments is that participation in the trial is a total experience (Syme 1978). An experiment may be designed to test the isolated effect to lowering blood pressure. But when subjects are instructed to take antihypertensive medications, and possibly to restrict salt and caloric intake and increase physical activity, they inevitably modify dietary fat intake, smoking, and other aspects of behavior.

The Multiple Risk Factor Intervention Trial (Kuller et al. 1980; MRFIT Group 1976a, 1976b, 1977; Sherwin, Sexton, and Dischinger 1979) recognized this limitation of single-factor trials. The protocol was designed to test the hypothesis that lowering serum cholesterol by diet, reducing high blood pressure by diet and drugs, and cessation of cigarette smoking, *in combination*, would result in a reduced risk of death from CHD. Men aged thirty-five to fifty-seven, who displayed various combinations of cigarette smoking, elevated blood pressure, and cholesterol, but who displayed no initial evidence of CHD, were to be followed for six years. After initial screening of 361,661 subjects during 1974–76, a total of 12,866 subjects were randomly assigned either to a program of special intervention (SI) directed toward these risk factors or to their usual source of medical care (UC). The experiment is being conducted at

MRFIT clinics in twenty-two sites across the country and is scheduled for completion in early 1982.

4.2.2 The Rand Health Insurance Study (HIS)

The responsiveness of medical-care demand to price is an important factor in the design of health insurance and the control of rising medical expenditures. Price elasticities of demand for medical services have been estimated from a variety of data sources. But the main source of price variation in these nonexperimental data is the terms of insurance coverage. Since consumers select their insurance on the basis of health status, income, family composition, and other factors affecting demand, such estimates could be seriously misleading.

The Rand Health Insurance Study (Manning et al. 1981; Manning, Newhouse, and Ware 1982; Morris 1979; Morris, Newhouse, and Archibald 1980; Newhouse 1974; Newhouse et al. 1979) was designed to over come this limitation. A sample of approximately 8,000 individuals in 2,823 families was enrolled in six sites across the country. Families were enrolled in one of fourteen different HIS insurance plans for either three or five years. These plans ranged from free care, to 95 percent coinsurance below a maximum dollar expenditure, to assignment in a prepaid group practice. Low-income families were over sampled. Persons eligible for Medicare, heads of households sixty-one years of age and older at the time of enrollment, members of the military, and the institutionalized population were excluded. Enrollment of subjects at the Dayton, Ohio site was completed in 1975, while enrollment at the Georgetown County, South Carolina site was completed in 1979. In addition to analysis of the effects of various insurance plans on medical care demand, the effects of coverage on health status (Brook et al. 1979; Ware et al. 1980), certain administrative aspects of health insurance, and the effects of HMO care are under study.

Both MRFIT and HIS can be legitimately called second-generation social experiments. Their designers took advantage of considerable prior experience in clinical trials and social experimentation. Nevertheless, these microexperiments exhibit important difficulties in design, execution, and interpretation. These difficulties will now be considered.

4.2.3 Subject Selection and Other Pre-Experimental Biases

In MRFIT, subjects were initially screened, primarily at work sites, by a series of medical examinations (Kuller et al. 1980). Those eligible at the first screening on the basis of blood pressure, cholesterol, and smoking habits were invited to a second, more detailed medical screening, at which time the purpose and duration of the study were explained. For those who returned for the third and final screening, informed consent was obtained and then randomization was performed. Since the trial was

aimed at men with high CHD risk, and since the experiment could not be blinded, potential subjects were necessarily informed of their medical status during the screening process.

It is reasonable to suspect that the initial volunteers in this experiment were highly motivated and therefore more susceptible to intervention than the general population. Of those subjects initially eligible by risk-factor criteria, about 30 percent declined to participate. Some of them merely refused to consider quitting smoking. It is also hard to imagine that the screening process itself had little effect on subjects' behavior and attitudes. Among those subjects who were ultimately randomized, mean diastolic blood pressure declined by about 10mm Hg from the first to the final screening examination, while the fraction of smokers declined by about 5 percent. Comparable changes were observed in blood cholesterol. These results may reflect changes in measurement methods between screening exams or statistical regression to the mean. Nevertheless, the evidence suggests that the pre-experimental phase constituted a form of life-style intervention.

The planners of MRFIT screened for subjects with high CHD risks in order to increase the statistical power of the experiment (MRFIT Group 1977).[2] But this practice is not without its problems. Blood pressure, cholesterol, and smoking are undoubtedly influenced by such factors as diet, stress, physical activity, socioeconomic status, family history, occupation, and peer pressure, many of which are difficult to measure. These additional, unmeasured variables also affect how subjects' CHD rates respond to experimental intervention. Pre-experimental screening on the basis of blood pressure, cholesterol, and smoking can produce a population of subjects that is highly unrepresentative with respect to the unmeasured variables. Some men who qualify for this study will be former quitters who have returned to the habit as a result of, say, transient job-related stress. Others will be light smokers who have transient elevations in blood pressure due to, say, excessive salt use or weight gain. Still others will be inveterate heavy smokers. Although the experiment would still yield an unbiased estimate of the effect of special intervention among those patients who qualified, it is not clear how the estimated experimental effect relates to the overall population response. This difficulty applies not only to experimental responses in risk factors, but also to the effect of intervention on CHD incidence. It is compounded further if the additional, unmeasured variables affect subject attrition during the experiment.

2. Selection was actually based on "modifiable risk," which is not necessarily synonymous with "high risk." This modifiable-risk score was based on a multiple logistic model of CHD risk, estimated from the Framingham study data (Truett, Cornfield, and Kannel 1967), in combination with educated guesses about differential success rates in reducing risk factors.

In the Health Insurance Study, the experimenters randomly sampled dwelling units and conducted initial interviews in order to ascertain the occupants' ages, incomes, and other data pertinent to eligibility. A base-line interview was administered to eligible families in order to elicit information about prior insurance status. Following verification of the insurance information, families were selected, assigned to the various plans, and contacted for an enrollment interview (Newhouse 1974; Morris 1979; Morris, Newhouse, and Archibald 1980). If the assigned plan represented less extensive insurance than the subjects had prior to entry, then the experimenters offered them a compensating incentive payment, in fixed installments, but unconditional upon subsequent medical-care consumption. Consent to participate in the study was elicited *after* these steps had been taken. Among families who completed base-line interviews and were assigned to treatments, 11 percent refused the enrollment interview. Of those who agreed to the enrollment interview, 27 percent refused the offer to enroll.

The HIS incentive payment scheme was intended to ensure that subjects in all treatment groups were no worse off financially by participating in the experiment. At worst, such payments were supposed to have a small income effect on demand. Nevertheless, with refusal rates in excess of 25 percent, it is worth inquiring whether prior assignment to a plan could have affected the decision to participate in the experiment. Those families assigned to the high coinsurance plans were more likely to receive incentive payments. In these families, the decision to participate should depend more heavily upon attitudes toward risk, expectations about subsequent health-care utilization, and other unmeasured variables. In fact, families who expect to make substantial use of medical care will be more likely to refuse to participate in the high coinsurance plans. It is at least arguable that these phenomena will result in an overly optimistic estimate of the effect of cost sharing on the medical-care use.

In both MRFIT and HIS, data have been collected on the characteristics of those subjects who refused to participate at the various pre-experimental stages, at least beyond the initial screening. It may thus be possible to assess some of the determinants of the decision to participate and to correct for potential nonparticipation biases. But the determinants of the decision to participate, it must be recognized, are not easily measured. So long as such intangibles play an important role, potential nonparticipation biases cannot be completely excluded. Moreover, replenishment of nonparticipants on the basis of observed characteristics, as suggested by Morris, Newhouse, and Archibald (1980), could be inappropriate.

4.2.4 Subject-Attrition Biases

Since MRFIT and HIS are still in progress, little information on attrition rates has been published. In the Health Insurance Study, the

three-year cumulative attrition rates for the free and nonfree plans have been 4 percent and 8 percent, respectively. In the MRFIT experiment, vital status has thus far been ascertainable for almost all of the participants. But the ascertainment of other morbid end points, such as nonfatal heart attacks, has been more difficult. Detection of these morbid events (by evidence on periodic electrocardiograms) required that subjects return for repeated checkups and examinations. At the end of the second year of the study, 6 percent of the special-intervention group and 7.2 percent of the usual-care group had missed their annual examinations. These proportions were 8 and 9 percent, respectively, by the fourth year. Among the SI participants, 16.3 percent had missed their biannual interim visits by the fourth year. The extent to which nonreporting subjects experienced a higher incidence of nonfatal morbid events is unclear.

It must be emphasized that subject attrition does not merely erode the statistical power of an experiment. Those who drop out may be least susceptible to the contemplated intervention. Certain imperfect covariates of the decision to drop out can be measured. But any attempt to correct for unmeasured determinants requires a model of the distribution of these determinants. The interpretation of the experimental effect may then be very sensitive to unverifiable assumptions about the parametric form of such a model (Harris 1982; Hausman and Wise, chap. 5 of this volume). In microexperiments, the only foolproof remedy for attrition bias is to keep subjects from dropping out altogether.

4.2.5 Hawthorne Effects and Anticipatory Responses

The subject's knowledge of his treatment assignment raises some serious problems for the MRFIT experiment. Although the usual-care subject does not receive the benefits of group sessions, counseling, behavioral therapy, and dietary instruction, he and his physician are informed of his risk status. Moreover, subjects in the UC group are asked, as in the SI group, to return for periodic visits and examinations. Highly motivated subjects who consent to randomization, but who end up in the UC group, may nevertheless alter their behavior. This phenomenon will reduce the contrast between UC and SI interventions and diminish the power of the experiment.

Preliminary reports from MRFIT (Sherwin, Sexton, and Dischinger 1979; Kuller et al. 1980; Schoenberger 1981) in fact show improvements in risk-factor scores for both SI and UC groups. After four years, SI men exhibited an 11 mm Hg drop in diastolic blood pressure, a 19 mg/dl drop in serum cholesterol, and a 41 percent smoking-cessation rate. UC men showed a 6 mm Hg drop in diastolic blood pressure, an 11 mg/dl drop in serum cholesterol, and a 23 percent smoking-cessation rate. Among SI men, 56 percent were being treated with antihypertensive drugs, compared to 41 percent in the UC group. These improvements could reflect further regression toward the mean or trends in behavior independent of

the experiment. But the motivating effect of the experiment itself can hardly be excluded.

MRFIT experimenters recognize that many years may be required before the observed changes in risk factors are manifested in reduced CHD rates. In that case, the long-term mortality results will hinge critically on subjects' behavior after the termination of formal life-style intervention. Perhaps the UC men, who received dramatic attention only in the pre-experimental period and who were forced to take responsibility for their behavior from the start, will display greater long-run improvements. By contrast, if SI subjects become dependent upon the experiment itself, then discontinuation of formal intervention could lead to higher relapse rates (Syme 1978).

The planners of the HIS have made special efforts to detect instrumentation artifacts and anticipatory responses (Newhouse et al. 1979). Participants' incentives to file insurance claims might depend on the amount of reimbursement. Hence, the plan assigment could affect subjects' reporting of medical-care utilization. To avoid this interaction between treatment and measurement of response, a system of weekly reminders to file claims was used. But the reminders themselves were also found to affect reporting. Therefore, a subexperiment involving biweekly probes was instituted. Since intrusive questionnaires and health reports could also affect subject desires to seek medical care, the sequence of examinations was similarly varied in a subexperiment. For the prepaid-care group, moreover, a set of "controls on controls" was employed, with no instrumentation at all. To ascertain whether certain subjects would earmark the incentive payments solely for medical care, the schedule of incentive payments and bonuses was also varied. In order to detect possible anticipatory responses to the beginning and end of the study, the experimenters plan to follow the three-year intervention group for an additional two years. They also plan to be watchful of initial declines in price elasticity after the onset of the experiment, followed by increases in price sensitivity as the end of the experiment approaches, followed by postexperimental responses to intraexperimental price changes (Arrow 1975).

It is difficult at this stage to see how all these instrumentation and anticipation artifacts can be estimated precisely. The issue here is not so much the separate, main effect of each form of instrumentation, but its interaction with treatment effects. There are too many interactions of instrumentation, treatment, and subject anticipation to test all of them satisfactorily. It is not completely clear how information on such artifacts can be easily incorporated into the final results.

4.2.6 Ethical Constraints

In the Multiple Risk Factor Intervention Trial, ethical considerations dictated that subjects with initial diastolic blood pressures above 114 mm

Hg be excluded from the study. Unfortunately, this form of sample truncation leads to difficulties similar to those encountered at the other end of the risk-factor scale. Thus, those individuals with previously undetected, severe hypertension may be derived from a population least motivated to seek routine care. These persons may have life-styles or other unmeasured characteristics that counteract or reduce any salutory effects of risk-factor reduction.

Even if a high-risk subject is eligible by screening criteria, ethical considerations dictate that treatment cannot be completely withheld. Hence, MRFIT does not compare treatment and nontreatment, but intensive intervention with "usual care." The usual care is not even average care, since the men randomized to the UC group have already undergone pre-experimental "treatment." Moreover, the planners of the experiment felt compelled to tell UC subjects that they were at high risk, including which risk factors were implicated (Kuller et al. 1980).

4.2.7 Interpretation of Treatment Effects

The design of MRFIT explicitly recognizes that people do not change their CHD risk factors one at a time. But its interpretation is still complicated by concomitant changes in dimensions of behavior other than the three risk factors. Subjects who are asked to change the saturated-fat content of their diet may also be influenced to increase their physical activity, which may in turn affect cardiac status. Men involved in a smoking-cessation group may alter their responses to stress, which could in turn affect cholesterol levels. Among SI subjects, in fact, nonsmokers and men who had quit smoking had the greatest improvements in serum cholesterol (Kuller et al. 1980, table 8). This makes it difficult to assess whether the effect of intervention resulted from changes in diet, serum-cholesterol levels, or other factors (Syme 1978). Furthermore, the methods of life-style intervention may vary considerably across the twenty-two clinical centers in MRFIT. Within a specific MRFIT clinic, treatments are further adapted to the idiosyncracies of the experimental subject. Even if we regard special intervention as a homogeneous entity, usual care remains ill defined. In the final analysis, if CHD rates improve with intervention in MRFIT, it may be difficult to know exactly what was responsible.

To be sure, one might attempt to elucidate the details of the experimental effect by specifying a response model. Thus, the Health Insurance Study was designed to estimate contrasts between the effects of different plans (e.g., the 95 percent coinsurance group versus the free-care group, or the prepaid-care group versus the remaining fee-for-service groups). But as early HIS data came in, the experimenters found the distribution of health-care expenditures to be highly asymmetric, with a discrete atom at zero expenditures and a fat right-hand tail (Manning et al. 1981; Manning, Newhouse, and Ware 1982). To perform statistical

tests of treatment effects, they therefore proposed a multiple-stage response model, involving the decision to seek care and expenditures conditional upon that decision. In addition to expenditures, health status was considered an important outcome measure. But health status could be both a determinant and a consequence of medical-care utilization (Brook et al. 1979; Ware et al. 1980). These considerations led the experimenters to some interesting, but even more complicated structural models of the experimental response. No doubt with further structural specifications, price elasticities and the parameters of response to deductibles and exclusions might also be estimated. I do not wish to denigrate these sophisticated efforts, but it should be pointed out that the conclusions derived from detailed-response surface models may be very sensitive to the structural specification assumed by the analyst. As discussed in several other papers in this volume, such models are far removed from the classical ideal of the one-way analysis of variance.

4.2.8 Relevance of the Results to Policy Options

Even if MRFIT clearly demonstrates a reduction in CHD risk, its special intervention does not necessarily correspond to a viable policy option. For one thing, widespread intervention at the individual level is expensive. Although employment-based health and fitness programs have become more prevalent, they may be quite different from the specialized research environments of the MRFIT clinical centers. Moreover, changes in life-style are likely to involve social learning, the diffusion of information, the changing of norms, and other phenomena that render individuals' responses interdependent. It is not clear that MRFIT captures these phenomena (Farquhar 1978; Kasl 1978; Syme 1978). Finally, such microexperiments reveal little about the effects of mobilizing voluntary health agencies, public restrictions on smoking, or the use of the mass media. Thus, MRFIT may reveal that CHD rates can be reversed. It may also offer some confirmation of the causal effects of risk factors. But it will offer much less information on the magnitudes of treatment effects in the general population. We could still be far from an operational public policy for preventing coronary heart disease.

The Health Insurance Study was designed primarily to be a *demand* experiment. Except for comparative analysis of responses at sites with different supply conditions, no attempt was made to assess the *supply* response to an insurance-induced increase in demand. Nor were the market-equilibrium effects of changes in coverage at issue. Yet the supply response to changes in insurance coverage is a critical factor in the recent rapid rise of health-care expenditures in this country (Feldstein 1977; Harris 1979, 1980; Newhouse 1978). Even after the HIS results are complete, policy makers contemplating changes in insurance coverage will still be uncertain about the effects of reimbursement on hospital

behavior, the consequences of insurance subsidy for technological change, or the effect of extensive insurance on competitive-market discipline.

The HIS, to be sure, focuses to a great extent on ambulatory-care demand. If the supply of ambulatory care were relatively elastic, and if the supply response of the ambulatory-care sector were independent of the remainder of the health-care sector, then the results of the experiment may offer a more complete picture of the ambulatory-care market response. Even so, the behavior of the elderly population, who consume a substantial and growing fraction of health-care costs, is not assessed in HIS. The decision to exclude the Medicare-eligible population from HIS was based on practical concerns about pre-experimental and experimental logistics. And a case can be made that experiment on elderly responses to insurance ought to be designed very differently. But if young and old demand from the same suppliers, then changes in the coverage of the under-sixty-five population could affect the price and access to care of the elderly. What is more, the redistributive effects of changes in insurance may be quite different in the market than within the confines of the microexperiment. At the very least, the proper application of the Health Insurance Study results to policy decisions necessitates the use of other nonexperimental data.

4.3 Possible Macroexperimental Remedies

I now set forth the background of an illustrative macroexperiment.

4.3.1 The Stanford Heart Disease Prevention Program (SHDPP)

From 1972 to 1975, the Stanford Heart Disease Prevention Program (Farquhar 1978; Farquhar et al. 1977; Meyer et al. 1980; Stern et al. 1976) conducted a field experiment in three California communities, each with a population of approximately 15,000. The objective of this pathbreaking study was to develop methods for modifying CHD risk that would be generally applicable to other community settings. Previous research had suggested that mass media campaigns directed at large populations could effectively transmit information, alter some attitudes, and produce small shifts in behavior such as influencing consumer product choice. But the effect of the media on more complex behavior was poorly characterized.

The planners of SHDPP therefore attempted a factorial experiment in which the combined effect of mass media and individualized intervention was assessed. From pre-experimentally surveyed populations in all three towns, they drew a subsample of men and women, aged thirty-five to fifty-nine, at high risk for CHD on the basis of cigarette smoking, blood pressure, and cholesterol level. In two towns (Watsonville and Gilroy,

Calif.), an extensive media campaign was conducted. In Watsonville only, two-thirds of the high-risk subjects were randomly assigned to individualized intervention, while the remaining third served as the media-only control. In the third town (Tracy, Calif.), no intervention was performed. Most of the reported results of this experiment have been derived from annual follow-up surveys of the original pre-experimental samples and the high-risk subsamples in the three towns.

Since the trial was to be coordinated from a single research center, intervention was restricted only to two towns. Although the assignment to individualized intervention in Watsonville was performed randomly, the allocation of media-based treatments was nonrandom. Although the three towns were geographically isolated, the overlapping television signals of Watsonville and Gilroy dictated that these two towns be assigned to media intervention.

4.3.2 Longitudinal versus Cross-Sectional Sampling

The planners of the SHDPP experiment relied upon longitudinal observations from a cohort of pre-experimentally screened subjects. Changes in CHD mortality statistics in each community over four years would have been too small to distinguish a treatment effect. Accordingly, a longitudinal sample may have appeared most appropriate to ascertain changes over time in behavior and knowledge of risk factors. Because media intervention was not randomly assigned, it may have seemed logical to use serial observations on many variables to bolster the claim that an observed effect was causal. But reliance on a cohort of pre-experimentally screened subjects leaves the experimental results wide open to many of the criticisms of microexperimentation, including selection artifacts, attrition biases, and Hawthorne effects.

Of the entire pre-experimental sample of 2,151 subjects in the three towns, only 1,204 actually completed all three follow-up surveys. The great fraction of those who failed to complete the study actively refused to participate or later dropped out (Stern et al. 1976, table 1; Maccoby et al. 1977, table 1). Among the 381 high-risk subjects who completed the baseline survey and who had not moved or died, 75 had dropped out after two years (Maccoby et al. 1977, table 2). By three years, the attrition rates among eligible high-risk subjects varied from 22 to 33 percent of eligible subjects across towns (Meyer et al. 1980, table 2). The average dietary cholesterol and saturated-fat intake, smoking prevalence and intensity, and systolic and diastolic blood pressures generally showed improvements over time in both experimental and control groups (Meyer et al. 1980, table 4). After three years, the only striking finding was that the subjects given both media exposure and individualized instruction had quit smoking at a higher rate than the other groups. Relative weight and blood pressure showed no difference, while the differential changes

in cholesterol were only suggestive. In view of these results, it is not unreasonable to suspect that the ultimate participants in SHDPP were highly motivated, that subject attrition was biased, favoring a positive treatment effect, and that many subjects were aware of the presence of an experiment.

These difficulties, however, should not be inherent to macroexperiments. Since the treatments are applied at the market or community level, there is no compelling reason why the responses in each unit should be obtained from a cohort. Sufficiently large, independent cross-sectional samples could be used to assess end points within each macro-unit. Since all of the residents in a community are subject to the same treatment, it matters little if different residents are sampled pre- and postexperimentally. Even in the case of certain morbid events of CHD, repeated cross-sectional samples of health-care providers could serve as a reasonable substitute for longitudinal samples. To be sure, these procedures sacrifice precision. But they avoid the biases engendered by subjects' decisions to participate and remain in a cohort, as well as their awareness of participation in an experiment.[3]

It is arguable that this trade-off between bias and precision does not differ from that encountered in microexperimentation. Thus, the experimenter who does not screen on risk factors or other dependent variables sacrifices statistical power. Overcoming this loss of precision requires more subjects, which in turn increases the cost of the experiment. However, the cost of increasing the size of repeated cross-sectional surveys within communities may be far less than the cost of including additional subjects in a longitudinal microexperiment, with all its follow-up interviews, diaries, and logs.

The advantage of repeated cross-sectional samples in macroexperiments is that individual subjects are less likely to be aware of the experiment. In fact it may be possible to perform blinded experiments, or at least blinded controls.[4] Even if some subjects became aware of experimentation, their incentives to avoid or anticipate the treatment may be weaker than in a microexperiment, where subjects can make decisions to participate separately from other economic choices. Thus, in a macroexperiment, an individual will have less incentive to leave a community merely to avoid certain media messages. So long as a different cross section is sampled on each round, refusals to respond are much less severe a problem. Of course, it is possible for an entire community to be aware of the presence of the experiment. But it is hardly clear that this is so undesirable. If the institution of an experimental policy causes antic-

3. In the Stanford Five-City Project, the Pawtucket Heart Health Program, and the Minnesota Heart Health Program, a mixture of cohort and cross-sectional sampling is being used.

4. A blind-control community is planned for the Pawtucket Heart Health Program.

ipatory emigration, or compensatory changes in local laws, or mass protests, that would appear to be a result worth knowing.

Repeated cross-sectional sampling in macroexperiments may further avoid ethical problems inherent in individual randomization. This is because the controls in a macroexperiment are "faceless," and the lives at stake are not specifically identified. To be sure, any subject found during sampling to be at high risk must still be informed of his condition and referred appropriately. However, so long as the experimenter samples from independent cross sections, and so long as the samples are not large in comparison to the population of the community, these ethical obligations should not materially affect the results. It is arguable that imposing involuntary participation on the citizens of a community is itself unethical (Hulley and Fortmann 1980), but I do not see this objection as insurmountable.

4.3.3 Costs of Macroexperimentation

Macroexperiments may incur lower costs of instrumentation, but the more difficult question is the costs of treatment. In a microexperiment, only those individuals who are recruited and sampled undergo treatment. In a macroexperiment, everyone in a community receives the treatment, even if his experimental response is not measured.

Certain types of macroexperiments, such as those involving price subsidies in large communities, are undoubtedly very expensive. But in many instances macroexperimental intervention may exhibit significant economies of scale. This applies especially to the use of mass media in SHDPP and related experiments, where the marginal cost of exposing an additional person to a health message is near zero.

4.3.4 Relevance of Macroexperimentation

Despite its problems of instrumentation, the SHDPP media experiment had one salient advantage over clinical trials such as MRFIT. The experimental treatment—that is, the use of mass media to transmit health information, to alter preferences, and possibly to change behavior—corresponded to a genuine policy option. The microexperiment may have revealed little about the social and behavioral mechanisms underlying the response to media intervention (Leventhal et al. 1980), but the elucidation of mechanisms should not be the objective of macroexperimentation. The main idea is to observe the effect of a contemplated policy in an experimental setting that closely approximates the environment in which the policy is to be applied.

The logical response, of course, is to ask whether the "black box" results of a macroexperiment are really relevant to the policy under consideration. Even if SHDPP and its progeny experiments should demonstrate an effect of media intervention on coronary risk factors and

rates, how do we know that media intervention will succeed in other communities? To this and related questions I now turn.

4.4 More Problems with Macroexperiments

4.4.1 The Confounding of Treatment Effects and Site Effects

My most serious concern about the Stanford three-city trial is the experimenters' assessment of the number of independent observations in their sample. In the early scientific reports on this study, the authors assumed that the number of independent observations equalled the total number of sampled subjects in the three communities. This assumption would be valid if applied only to the Watsonville microexperiment in which subjects were individually randomized. But for the mass media macroexperiments, there were really only three independent observations.

Confusion over the number of degrees of freedom in macroexperiments has been widespread. In fact, the issue appears to have been resolved, broached all over again, and then settled several times in the literature. Yet biostatisticians continue to propose formulas for appropriate sample size in community trials as if the individual were the unit of randomization (Gillum, Williams, and Sondik 1980).

The confusion derives in part from the view that outcome measurement in community-prevention trials is merely a form of cluster sampling (Cornfield 1978; Gillum, Williams, and Sondik 1980). If the experimenter wishes to estimate, say, CHD death rates, then sampling by community, rather than by individuals, will increase the variance of estimated population rates. The increase in variance would be inversely related to the degree of homogeneity of death rates within communities and directly related to the extent of heterogeneity between communities. Hence, if the experimenter could select relatively homogeneous intervention sites, the loss of efficiency would appear to be minimal. But this view ignores the fact that an experiment has been conducted and must be interpreted. The real issue is that in the interpretation of the results, the "site effects" are confounded with the "treatment effects."

Consider the following example. Suppose that community A is chosen for a media campaign and community B is selected as control. Suppose further that we could randomly allocate N subjects each to live in these two towns. Each subject, it is assumed, belongs to a homogeneous population with respect to pre-experimental risk of CHD. How should we interpret the results of the media campaign? If we believed that the two communities were merely artificial vessels for separating experimental from control groups and that within each community there

was no intercorrelation of subject responses, then we have 2N observations on the two treatments. But if the billboard density in a community affects the frequency of messages, or the ideology of the local television station owner affects the prominence of health-related commercials, or if the configuration of voluntary agencies affects opinion leadership, or if social networks permit greater diffusion of information, or if subjects' responses depend on their conformity with others, or if subjects' changes in dietary habits depend on food prices in a community, then we no longer have 2N independent observations. Even if we could randomly assign subjects to communities A and B, the results could be quite different if town B were instead chosen for intervention and town A were instead chosen for the control. Moreover, it would not help to assess the pre-experimental variance of death rates between and within communities. By construction, these variances would all be zero. The issue is not pre-experimental death rates, but the responses of death rates to the intervention.

To be sure, site effects are common in microexperiments, such as MRFIT, where the size of the experiment dictates the deployment of multiple clinical centers. But the situation in microexperiments is considerably different because randomization of subjects takes place within each site. Hence, site effects can be distinguished from treatment effects, and site-treatment interactions can be tested.

The literature on clinical trials is replete with tests of site effects and site-treatment interactions (e.g., hospital effects in the National Halothane Study, clinical-center effects in the University Group Diabetes Program trial of insulin or oral hypoglycemic agents versus placebo). Hopefully, in the analysis of the final results of MRFIT, treatment successes at particular clinical centers will receive scrutiny. But in pure macroexperiments, there is no crossover of treatments within a community. The site effects are fully nested within the treatments. Sampling more subjects at each site will diminish the variance of the estimated death rate within each site, but it will not affect the precision of these site-treatment interactions. In fact, if we have only two treatments and two sites, there are no degrees of freedom to disentangle these treatment-site interactions. Only more sites will solve this difficulty.

4.4.2 External Validity

When the experimenter tests for site-treatment interactions, he is asking whether any specific characteristic of a market or community could be uniquely responsible for, say, an observed effect of media campaigns. If he samples enough communities, he can distinguish between a general media effect, applicable to all sites, and media effects that are merely idiosyncratic for certain communities. But then how does the experimenter know that the selected sites constitute a representative

sample of these idiosyncracies? What would be the effect of media intervention in communities where a single, large employer also started his own employee health program, or where a national manufacturer test marketed a new, low-cholesterol product? If relatively small towns were selected, as in SHDPP, what would the results tell us about the effects of intervention in large cities? Would they be relevant to macroexperiments on work groups or domiciliary institutions (Rose et al. 1980; Sherwin 1978; WHO European Collaborative Group 1974)?

So long as the site-treatment interactions are regarded as random effects, the experimenter is obligated to choose judiciously experimental sites that are representative of the environment in which the policy is to be instituted. I recognize that even in macroexperiments, one ought to select sites that are not wholly unrepresentative. It is thus worth inquiring whether the communities selected for HIS possess doctors, hospitals, medical standards, and institutions that are typical of the United States. And I have already inquired whether the clinical centers in MRFIT are representative of programs of individualized intervention throughout the country. But it seems to me that the burden on macroexperiments is much greater.

4.4.3 Randomization of Macrounits

Many of the proponents of community-based intervention trials regard randomization as an impractical ideal. There are just too many administrative political obstacles. Unfortunately, I see virtually no way out of the requirement that experimental sites, once selected, must be allocated randomly to treatments. I acknowledge numerous instances where evidence from nonrandomized studies has proved convincing. But in those cases, the analysis has hinged on a paucity of plausible rival explanations for the observed difference between treatment and control groups (Campbell and Stanley 1966). But in macroexperimentation, there is likely to be an abundance of rival explanations. It is not hard to imagine that a town with its own television station or health-conscious opinion leaders will be more willing to undergo a media campaign. Such a community may be more susceptible to the effects of such an intervention.

4.5 Toward a Science of Macroexperimentation

Despite substantial advances in design, execution, and interpretation, microexperiments still have serious and possibly inherent difficulties. Individuals make nonrandom decisions to participate or drop out of the experiment. They may be influenced by the instrumentation process. Even in the absence of these difficulties, microexperiments do not necessarily test real policy options. Macroexperimentation, on the other hand,

may avoid some of these problems. But convincing macroexperiments require many observations at the community or market level. Moreover, political and administrative factors may dictate nonrandom selection of communities, with its attendant difficulties. And there is always uncertainty whether the observed effect of treatment in a sample of communities was not due to idiosyncratic, unrepresentative characteristics of the experimental sites.

We are thus faced with a serious dilemma. Should we perform a microexperiment, optimistic that instrumentation artifacts will not arise, and thankful to learn something about one aspect of a complicated policy problem? Or should we plunge ahead with a "sloppy" macroexperiment, with all of its difficulties of interpretation and generalization?

4.5.1 Decentralized Macroexperiments

Because SHDPP was to be coordinated by a single research center, the experiment was restricted to only three towns. Once these three were selected, random assignment to media exposure was made impossible by overlapping television signals. But it is worth speculating what experimental design might have arisen from a multi-center trial. If the Stanford group had been one of many research centers, couldn't they have selected a pair of towns, both of which had nonoverlapping television signals? Why couldn't treatment be randomly assigned between the two towns? Why couldn't the Stanford city-pair be one block in a larger matched-pair experiment?

My point here is that many of the most serious difficulties of macroexperiments may result from overcentralization. So long as we could allocate pairs of comparable sites (or perhaps larger subsets) to individual experimental blocks, the execution of each block could be the responsibility of a separate research center. Within each block, randomization may be more feasible. Increasing the statistical power of the experiment, and perhaps its external validity, means increasing the number of blocks.

Such a design is not entirely speculative. In fact, the WHO European Collaborative Group (1974; Rose et al. 1980) has been conducting a macroexperiment in CHD prevention in twelve pairs of factories in various cities. These factories (or in some cases occupational units within factories) were recruited into the trial before random assignment to treatment or control. The factory pairs were matched as far as possible by age, geographical area, and the nature of the industry. The subjects include all male employees aged forty to fifty-nine years, not merely those at high risk. This design unfortunately involves longitudinal follow-up of cohorts. Hence, it may be susceptible to participation biases, selective employee turnover, and Hawthorne effects. But it illustrates the possibility of randomization within blocked pairs of macro-units.

One might object that only small units, such as factories and domiciliary institutions, are susceptible to randomization (Sherwin 1978). Larger political entities will merely balk at the uncertain prospect of receiving the less desirable assignment. But it is hardly clear to me that this state of affairs is inevitable. For one thing, the possibility of randomization among matched pairs may be more palatable politically than random drawings from a larger population of sites. In some cases where the eligible sites are political subdivisions under the governance of a higher authority, the possibility of site self-selection may not be so serious. In fact, several macroexperiments in cancer screening, in which census tracts, townships, or counties are the relevant sites, have already been proposed (Apostolides and Henderson 1977). Moreover, in cases where communities or organizations have already received some type of government grant or benefit, the continued receipt of that benefit could be made the incentive for participation in the experiment. In cases where various communities apply for grants to become demonstration sites for a particular innovation, the awards process could be broken down into two stages. A subset of deserving, eligible sites would first be chosen. Among eligible sites, treatment and control assignments could then be made. It is remarkable to me how often government agencies and other grantors first make the awards to the most deserving sites and then ponder how a comparable set of control sites is to be chosen from the losers for the purpose of project evaluation.

When intervention at a large number of sites is managed by one research or administrative group, the inevitable consequence is a rationing of limited intervention effort to a few sites. In extreme cases, many of the so-called intervention sites do not receive any intervention because the research team has merely lost control of the project. Administrative decentralization of macroexperiments could allay some of these problems. Moreover, some degree of blinding may be possible. At the least, a research team responsible for intervention in one block of sites need not know the progress of the experiment in other blocks.

4.5.2 Time-Series Experiments and Crossover Designs

The possibility that communities or other macro-units could serve as their own controls has not been adequately explored. Admittedly, any comparison over time is susceptible to confounding interpretations. Experimental responses take some time to be completed. What appears to be the effect of a cross-over may actually be a transient from earlier intervention (Morris, Newhouse, and Archibald 1980). If the macroexperiment is not blinded, then the effects of crossover could be confused with anticipatory responses or other Hawthorne effects. Nevertheless, there is a variety of familiar devices for detecting time-varying responses.

Although these devices have been derived from microexperiments, they could at least be tried in the macrosetting.

For example, in the case of a matched-pair design, the treatment and control communities could reverse their assignments later in the experiment. The timing of this reversal need not be scheduled in advance, or at least known to the experimental units. Stopping short of complete crossover, I could also envisage folding-back designs. We could begin by a series of observations on communities in which no intervention is instituted. Thereafter, one or more of the communities becomes a treatment site. In sequence, the remaining communities receive the intervention. Again, the sequence and schedule of assignment could be random and unknown to the experimental units. If all of the units are destined ultimately to receive the intervention, randomization with respect to the sequence and timing of the intervention may not present so many political or administrative obstacles. Such folding back designs may be particularly useful when the endpoints are subject to habit formation and thus difficult to change in short intervention intervals.

4.5.3 Mixed Macro and Micro Designs

In some cases, a mixture of micro and macro designs might enhance the power of the experiment. Such cases arise when the interventions at the individual and site levels are qualitatively similar.

In the SHDPP trial, a subexperiment of individual intervention was performed within Watsonville, a town receiving media intervention. This subexperiment was designed to test the interaction between the two types of experimental treatments. Unfortunately, the investigators were unable to conduct an identical subexperiment in Tracy, the town receiving no media intervention. But even if a full factorial design had been undertaken, the two types of treatment were qualitatively different, so that only their crude interaction could be profitably investigated.

In other cases, however, both interventions could be close enough to conform to a simple response model. Suppose, for example, that the experimenter wishes to investigate the effects of varying employer contributions to employee health-insurance premiums. Since changes in employee benefits are typically performed at the level of the firm, a macroexperiment would be appropriate, with various firms corresponding to different macro sites. But within each firm, employer contributions could be further varied among employees. Such an experiment could offer considerable insight into firm-specific and employee-specific responses to changes in employee premium subsidies.

4.5.4 Combining Macroexperiments

A potential significant advantage of macroexperimental blocking is its ability to enhance the external validity of the experiment. Within each

block, experimental sites might possess similar characteristics, but between blocks the site characteristics could vary considerably. In community-based life-style intervention, it would be especially informative for blocks to vary with respect to the size, climate, age structure, sex, racial, and ethnic composition of their member communities.

A number of independent community-based life-style-intervention trials are already in progress in this country. Taken together, these trials might be considered a single macroexperiment with multiple blocks. The difficulty with this interpretation, however, is that the method of intervention may vary considerably from one block to the next. We thus cannot easily distinguish between a block effect and a block-treatment interaction. If some community trials show significant effects of life-style intervention and others do not, it will be unclear whether the discrepancies resulted from differences in the type of media intervention across trials, or differences in the susceptibility of communities to media messages. The results of different trials could be combined only if we had some prior information on the relationship between types of media intervention employed.

Some recent theoretical work on combining diverse experiments might be usefully applied to this problem (DuMouchel and Harris 1983). A complete exposition is necessarily beyond the scope of the present paper. But the main idea is to specify formally a structural relationship between the treatment effects in each community trial. For example, the magnitude of the effect on CHD rates might depend on the extent of electronic-media intervention, the duration of intervention, or the recruitment of voluntary agencies. A model of the treatment effect that relates these characteristics is then superimposed upon the results of each trial. The main issue in the application of such a technique is the degree to which life-style intervention in each trial was independent of the characteristics of the communities under observation. For example, if the experimenters in a particular trial resorted to scientifically oriented media messages because the target communities were highly educated, it may be impossible to distinguish between the treatment effect of media content and the role of educational background in a community's response.

4.5.5 Competition Experiments, Regulation Experiments, and Deregulation Experiments

Reduction of the tax subsidy on health-insurance coverage, elimination of barriers to entry for prepaid health-care providers, and enhancement of consumer choice of health-insurance plans have been proposed to control rising health-care expenditures. Virtually all of the evidence supporting the efficacy of these interventions is nonexperimental. Our policy makers could, of course, take the available data as sufficient cause to plunge ahead with a full-scale policy. But the correct course, it seems

to me, is to assess some of these innovations experimentally before taking such a precipitous step. I have already hinted how several large employers in a number of different cities might serve as sites for experimental changes in employee health-insurance benefits. Perhaps several distinct divisions of the same large corporation could form an experimental block. Community-based experiments, in which the effects on market competition are observed, are also conceivable.

Regulatory controls on health-care expenditures have also been suggested. Although various innovative forms of hospital reimbursement have been tried, most of the so-called reimbursement experiments have really been uncontrolled demonstration projects. In view of the substantial likelihood that hospitals subject to those novel controls have been selected in a biased manner, it is hard to know exactly what significance these projects should have for future policy decisions. It is difficult for me to see why the experimenters have not blocked participating hospitals according to, say, size, teaching status, or range of facilities, and then randomly assigned the novel form of reimbursement within each block.

One variant of the fold-back design discussed above is the deregulation experiment. In this case, the experimental treatment is the removal of an intervention already in place. The sequence and timing of deregulation at various sites is the critical control variable. This type of design may be particularly useful when the value of a regulatory program is in question. Even if our policy makers deem that physician-peer review schemes or health-planning agencies are to be discontinued, it would be valuable to learn something about the effects of these policies during their demise.

4.6 Conclusions

This paper can be easily criticized for its lack of balance. I have sought out the most subtle crack in microexperiments, yet I am willing to cover large faults in macroexperiments with hopeful speculation.

The plain truth is that macroexperiments in public policy—or at least corrupted versions of macroexperiments—are far more prevalent than the microexperiments to which social scientists have devoted so much attention. It is not too soon to develop some meaningful strategies for effective macroexperimentation.

4.7 Epilogue, 1981–84

After this paper was written (spring 1981), the main results of the Multiple Risk Factor Intervention Trial were published (MRFIT Group 1982). Although CHD deaths in the special-intervention group were 7 percent less than in the usual-care group, the difference was not statistically significant. One reason for the weak results was the unexpectedly

low death rate of the usual-care group—40 percent lower than expected. The usual-care subjects apparently benefitted from the information about their high CHD risk and the provision to their physicians of original and follow-up medical data.

Although both the experimental and control groups showed declines in blood pressure, cholesterol and cigarette use, nevertheless the experiment was singularly successful in achieving much greater smoking cessation in the SI group than in the UC group (figure 1 and table 2 in MRFIT Group 1982; also, Ockene et al. 1982). Among men who were smokers at initial screening, however, mortality differences between the SI and UC groups were modest (table 5 in MRFIT Group 1982). Yet among all subjects (both SI and UC) who smoked at the time of entry, those persons known to have quit smoking in the first study year had considerably lower subsequent death rates than those known to have continued smoking (table 9 in MRFIT Group 1982). A plausible interpretation is that smokers who missed their first year follow-up visit had much higher subsequent death rates than those smokers who reported their status. Moreover, the mortality differential between those who missed follow-up visits and those who returned was more marked for the SI subjects. Thus, special intervention was apparently more effective than usual care in producing attrition among the really sick people. Those SI subjects who remained in the intervention program had lower CHD mortality, but not much lower than those nonattriters in the UC group.

After the current paper was written, the Stanford Heart Disease Prevention Project published a reanalysis of the three-city data (Williams et al. 1981). The new analysis acknowledged that the communities, and not the individual subjects, could be the experimental units. For such endpoints as cholesterol and blood pressure, the authors computed mean values for each of the three towns and for each of the four years of the study. The slopes derived from linear trend regressions on each town were then compared. By stacking together the slope estimations in a single regression, the authors were able to make a few statistically significant inferences, but the results still had far less precision than those previously reported.

After this paper was written, the Rand group published a number of "interim results" of its Health Insurance Study. Such results were based on about 40 percent of the total person years that are ultimately available for analysis (Keeler et al. 1982; Newhouse et al. 1982; Duan et al. 1983). In comparison to free care, copayment for medical services was found to reduce the number of ambulatory visits and the number of hospitalizations among adults, but not the cost per hospital stay. When health-care use was aggregated into episodes of illness, copayment was found to reduce the number of episodes but not the cost per episode.

The interpretation of the Rand findings is not obvious. More than

two-thirds of hospitalized subjects incurred expenses that exceeded the maximum expenditure for even the highest copayment plans. Hence, most hospitalized patients faced marginal coinsurance rates that were effectively zero. An alternative explanation is that patients with full coverage were hospitalized with less serious and thus less costly illnesses. Or perhaps patients have little or no influence on the disposition of care once they have sought treatment. In any case, these findings highlight the study's limited focus on the demand side of the medical-cost problem. The continuing rise in medical expenditures reflects increases in the costs per hospital stay and no doubt the costs per episode of illness. On the demand side, these critical variables may to be unaffected by realistic changes in coverage. But what about the supply side?

Statistical analysis of the HIS results has not been so simple. The pattern of health-care expenses for each of the plans included a substantial fraction with zero claims. The distribution of positive expenditures showed a long right-hand tail caused by rare, very large claims. Thus, confidence intervals derived from the conventional normality assumption were quite large. To improve precision, the Rand investigators devised a four-equation regression model to assess the effects of the experimental plans. In the first probit equation, the probability of medical use depended on plan dummy variables and various covariates (for example, physician visits predicted from 1971 national data based on the age and sex of each subject). In a second probit equation, the probability of hospital expenditures conditional on use of care depended on additional interaction effects between age and plan that the Rand authors discovered to be important. In a third regression equation, the logarithm of expenditures among those with only outpatient use had a variance component for intrafamily effects. In the fourth regression equation, the logarithm of expenditures among those with inpatient care did not depend on dummy variables for individual plans. Because of large outliers, the latter equation was estimated by a robust weighted regression method. To correct for the bias in transforming the predicted mean expenditures from the log scale back to the arithmetic scale, a new nonparametric estimate was developed. The standard errors for the transformed means were then estimated from first-order approximations (Duan et al. 1983). The authors have acknowledged that predicted expenditures by plan (and their confidence intervals) are highly model-dependent and that there is the danger of overfitting the data. Much to their credit, they have performed some interesting tests for such overfitting on a subsample of the interim data. But they do concede with appropriate caution that later analysis of the full experiment may lead to further modeling changes as more data are accumulated at the far right tail of the expenditure distribution.

In retrospect, my paper glossed over certain problems of macroexperimentation that deserved more careful scrunity.

First, I did not address what types of policy interventions are accessible to macroexperimental analysis. The media campaigns of SHDPP were obviously suited to community-wide study. But many policy interventions are aimed at small, diffusely scattered populations of eligible persons. How, for example, might we assess a proposed plan for insurance coverage of a particular medical intervention such as organ transplantation or hospice care for terminal illness? Here, we need to be more creative in defining appropriate macroexperimental units, such as transplantation centers or individual hospices.

Second, I merely suggested without strong supporting evidence that macroexperimentation might be more immune to the Hawthorne effects, selection and attrition biases, and other artifacts that have plagued microstudies. Certainly, if we sampled towns with pre-experimentally high childhood leukemia rates and then in half of them compelled residents to drink only bottled water, we might very well see leukemia rates fall (by regression to the mean) or maybe a large confounding population exodus. But for more realistic cases, it is a serious empirical question whether such artifacts will be significant. In MRFIT, to be sure, we might know enough about intertemporal variation in an individual's serum cholesterol levels to correct for potential regression to the mean and pre-experimental selection bias. But it is not obvious that comparable data on intertemporal variation in site characteristics would be so scarce.

Third, I acknowledge that repeated, independent cross sections would result in much less precise intrasite estimates than might be afforded by cohort sampling. But I avoided asking exactly how much precision might be lost by the use of such cross sections. Certainly, if the endpoint under consideration displayed extremely high intertemporal correlations among individuals, the required sample sizes might be an order of magnitude larger. But it is not obvious that the cost of such cross-sectional sampling will be so much larger.

Fourth, I was too cavalier about the generalizability of macroexperimental results. The success of a macroexperimental study of the use of electronic media in health promotion might depend, say, upon which celebrities gave testimonials. The effects of changes in tax treatment of health insurance among various experimental corporate sites might depend, say, upon the relations between organized labor and top management. The effects of alteration of physician payment for hospital-based care might depend, say, upon the facilities available at the site hospitals. Such uncertainties are inherent in any form of public policy evaluation. Macroexperimentation, however, may be better equipped to overcome such challenges to external validity.

Fifth, I did not meet the challenge of designing a macroexperiment analogous to the HIS. The difficulty I encountered here is that a health-insurance macroexperiment would end up asking questions quite different from those asked in the Rand study. Do changes in insurance coverage affect the rate of introduction of new techniques into a market, or the rate of entry of hospitals, prepaid plans or other providers? Would expansion of coverage result in various health-care rationing schemes, including queues, triage, or more regulation? To be sure, observed changes in entry into experimental communities (from nonexperimental areas) might not mimic the responses to a nationally available insurance system. It may take considerably longer for suppliers' responses to changes in insurance to reach long-run equilibrium. Still, the questions are too important to be ignored.

Finally, wasn't I just kidding myself about the real costs of macroexperimentation? Wouldn't large-scale interventions entail enormous administrative and treatment expenses? I think not. We are constantly instituting new demonstration projects and innovations in the health-care arena without careful advanced planning as to the ultimate evaluation of such efforts. The genuine costs of macroexperiments lie in the additional resources required to look forward as well as back.

Comment Paul B. Ginsburg

Jeffrey Harris's stimulating paper argues that we have had an imbalance between social microexperiments and social macroexperiments. Drawing upon the experience of experimentation in the health area, he shows that microexperiments have had serious problems that would be difficult to correct, while the problems with macroexperiments tend to be more amenable to solution through clever experimental design.

The paper describes clearly the seriousness of some of the following obstacles to the validity of microexperiments: 1) biases in the selection of subjects and attrition, 2) anticipatory responses and Hawthorne effects, 3) ethical restraints on randomization, and 4) interdependence among individuals.

It then discusses how macroexperiments can avoid these problems. For example, macroexperiments can study market equilibria, thus recording the effects of interdependencies among individuals. By not requiring individual volunteers, selection biases are eliminated. The nature of intervention in macroexperiments also avoids many ethical constraints,

Paul B. Ginsburg is a deputy assistant director, Human Resources Division, Congressional Budget Office.

such as the need to inform control-group participants of the presence of medical conditions and the value of conventional treatments.

Nevertheless, macroexperiments do have some serious disadvantages. One is reduced statistical power. Harris points out that the relevant number of observations in a macroexperiment is the number of sites, not the size of the affected population. Given the inability to control other determinants of the outcome in question, inferences from a handful of sites have limited statistical power. Another problem is the administrative and political obstacles to randomization.

Harris's paper is a valuable one. His critiques of microexperiments are clearly presented and convincing. Rather than simply listing theoretical problems, he makes a careful case about their importance for validity. His ideas for overcoming some of the problems in macroexperiments are good ones that will benefit social experiments.

While I agree with many of the points that Harris makes, I am somewhat uncomfortable with his characterization of the choice as one between a microexperiment or a macroexperiment. I wonder how frequently both options are practically available and are the first and second choices. A more common choice is between an experiment and collection of nonexperimental data. With the Health Insurance Study, for example, I would expect those most critical of the problems encountered by it to advocate increased collection of nonexperimental data rather than a macroexperiment with national health insurance. The latter would be quite expensive, and its limited time period would lead us to question whether full-fledged market effects are being observed. Indeed, the nonexperimental alternative to the Rand Experiment was actually performed, funded by a different agency in the Department of Health and Human Services. The National Medical Care Expenditure Survey, sponsored by the National Center for Health Services Research and the National Center for Health Statistics took a substantial step forward from previous health-care surveys by employing periodic interviews and obtaining direct information from insurers, employers, and medical providers to supplement that obtained from the respondent.

Often the choice of micro- versus macroexperiments is dictated by the nature of the proposed intervention. Since SHDPP used mass media as the intervention, no choice between a micro- or macroexperiment existed. When an intervention specific to individuals is the object of study, there is a theoretical choice, but expense often renders the macro version unrealistic.

I am in agreement with Harris that creativity on the part of researchers can yield a great deal of macroexperimental analysis. Government is frequently initiating (and more currently terminating) programs. Budgetary, administrative, and political constraints often require that programs be phased in or phased out. Participation in the design of this process by

researchers could tremendously increase its potential generation of evaluative information.

One program of this sort that I am familiar with is the Professional Standards Review Organization (PSRO) program, which reviews the appropriateness of medical services delivered to Medicare and Medicaid patients. The program was phased in by funding as many local volunteer organizations as the federal budget would permit. While the willingness of a local physicians' group to participate was important to the workings of the program, randomization among the volunteers could have been performed, even if it resulted in some delay in implementation.

Now the program is being phased out. The Department of Health and Human Services is selecting for defunding those agencies it feels are least effective. Since ability to distinguish between those more effective than others is limited, a larger list of the least effective organizations could be developed and randomization performed on this list to choose which ones to defund.

Harris's idea for getting information from demonstrations is a good one. He is correct that the manner in which organizations are chosen for demonstrations prevents useful inferences, but that randomization among the volunteers could provide meaningful information.

The suggestions concerning social experiments for competition in the financing and delivery of health care are interesting. A true macroexperiment, involving selecting certain markets for changes in tax policies and changes in Medicare reimbursement, is probably not feasible. But an experiment would be feasible and useful to test employees' responses to a choice of insurance plans, which is perhaps the link in the competition model that the least is known about. The experimenter could probably even simulate tax-free rebates without changing the tax law by making payments to offset taxes due. I do not know whether such an experiment would be characterized as micro or macro. Clearly it has elements of both. Its results would be far more useful than those reported by employers initiating such programs on their own.

Comment

Lawrence L. Orr

Choosing between Macroexperiments and Microexperiments

Jeffrey Harris argues that "economists and other social scientists . . . have spent disproportionately too much effort on the design and interpretation

Lawrence L. Orr is director, Office of Technical Analysis, A.S.P.E.R., U.S. Department of Labor.

of microexperiments" and suggests that greater attention should be given to the potential use of macroexperiments. He defines microexperiments as those in which "the experimenter assigns treatments and gauges responses at the level of the individual," whereas "in social macroexperiments, treatments are assigned at the group, community, or market level." While Harris is careful to assure us that "this paper is not a broad endorsement of macroexperiments" and "does not advocate the abandonment of microexperiments," the theme of the paper is that microexperiments are subject to a long list of inherent defects that, one gathers, render confident interpretation of the results almost impossible, whereas the (much shorter list of) shortcomings of macroexperiments are remediable through clever design.

On the basis of my own experience with both types of experimental research,[1] I find both Harris's indictment of microexperiments and his enthusiasm for macroexperiments seriously overdrawn. Perhaps more fundamentally, I think that he has not posed the central question in the most useful way: The real question is not which type of experiment is "better" in some absolute sense, but which is more appropriate to the problem at hand.

Harris and I appear to have fundamentally different views of the role of experiments in the policy process. Harris appears to take as his starting point a single well-defined program (or, at most, a few) of unknown efficacy; the role of the experiment is to provide a comprehensive, holistic evaluation of this program or programs, so that the policy maker can make a simple go/no go, adopt/reject decision.

This approach leads him naturally to what I would term "black box" experiments, applied to whole populations with or without experimental variations, relatively simple aggregate-outcome measures, and little or no analysis of underlying response behavior. In contrast, I tend to assume that the policy maker starts with a whole range of program options that can be characterized by a finite set of policy parameters (tax rates, subsidy levels, staff/client ratios, etc.). The function of the experiment, then, is to provide measures of the response to these policy parameters that will enable the policy maker to select that combination, or those levels, of policy instruments that achieve the "best" outcomes, i.e., to design the program. This paradigm leads me naturally to experiments with many variations and extensive analysis of micro data in order to estimate individual response functions.

1. In recent years, I have had some involvement in the design, execution, and/or analysis of the four income-maintenance experiments conducted by OEO and HEW, HUD's housing-allowance experiments, the OEO/HEW health-insurance experiment, HEW's experiments in AFDC administration and disability insurance, and DOL's Employment Opportunity Pilot Projects and (the stillborn) Positive Adjustment Assistance Demonstrations. All except the DOL projects were (primarily) microexperiments; the DOL projects fit Harris's definition of macroexperiments.

This latter view of the role of experiments seems to me in keeping with the way we treat most other research—we seldom expect individual nonexperimental research projects to render global assessments of major policy initiatives—but I will concede that there is a nontrivial set of policy questions for which the black box experiment is appropriate. As I have already suggested, the trick is to figure out which policy issues are in that set.

To do that, though, one must have a full appreciation for the relative strengths and weaknesses of the two modes of experimentation. Therefore, in what follows, I will first discuss briefly the methodological issues raised in the paper with respect to microexperiments and some of the problems of doing macroexperiments, before attempting to lay out a general set of criteria for choosing between the two in addressing any particular policy questions.

I should note at the outset that much of my experience with experimentation is in nonhealth areas and that therefore many of the examples and counterexamples in what follows relate to nonhealth interventions. The issues raised by Harris, however, are primarily methodological ones that cut across substantive research areas, so the actual subject matter under investigation is often of secondary importance to the argument. If one is to argue from example—and it appears that in many cases that is the best we can do at this stage of development of the art of experimentation—it seems to me that more examples are preferable to less.

There is no question that a number of serious methodological problems are encountered in designing and interpreting microexperiments; most of them are listed in this paper. But it should be recognized that many of these problems are not peculiar to microexperimentation; they apply with equal force to many other types of empirical research, including macroexperiments. Thus, for example, the problems of misreporting, interview refusal, attrition, and Hawthorne effects are really problems of longitudinal survey research. Any researcher doing nonexperimental statistical analysis of the Current Population Survey or the Health Interview Survey faces these same problems, although it is my observation that nonexperimental researchers are much less likely than experimenters to recognize or attempt to deal with them. Moreover, these problems will also afflict any macroexperiment that relies on surveys for its data base (as Harris himself notes).

Some of the problems posed by Harris—for example, the necessity of using "detailed-response surface models" for analysis and the constraints imposed by ethical considerations—are inherent in the problem being addressed, not in the experimental methodology. It seems self-evident that sorting out the causal relationships among health-insurance coverage, consumption of medical care, and health status is an exceedingly complex endeavor that is likely to require complex analytical models,

however the research data is generated. Likewise, if it is unethical to provide (or withhold) a particular treatment to a randomly selected individual in a microexperiment, it is hard for me to conceive that it is ethical to do so to a group or entire community in a macroexperiment.

There are, of course, limitations that are inherent in microexperimentation itself. Microexperiments are inevitably of relatively short duration and therefore may not reveal long-run, steady-state responses. This characteristic is, of course, shared by macroexperiments. If the likelihood of bias in a particular application seems serious, the researcher might be well advised to consider some nonexperimental data source, such as observations on "natural experiments" or data from ongoing programs, instead of—or in addition to—experimentation. Likewise, so long as participation in microexperiments is voluntary, selection bias is an ever-present danger. As I argue below, however, closely analogous problems exist in macroexperiments. Selection bias is, of course, endemic in nonexperimental data.

Perhaps the most fundamental criticism raised by Harris—and the point on which we disagree most strongly—is the relevance of microexperiments to policy. In the context of the Multiple Risk Factor Intervention Trial, he argues that the treatment "does not necessarily correspond to a viable policy option." This is so, he argues, because intervention at the individual level is expensive; public policies are more likely to take the form of organizational, educational, or regulatory efforts aimed at diffusing information or changing behavior in the community at large. I certainly agree that MRFIT will not predict the outcome of those policies. If the objective was to learn what effect, say, a particular mass media educational campaign would have on aggregate rates of coronary heart disease in the community, then by all means that is the policy that should have been tested, and it could probably only be tested with a macroexperiment. The fact that the researchers did not do so indicates that either that was not their objective or that they showed poor judgment in their choice of research strategy; it does not strike me as an indictment of microexperimentation per se, beyond the rather obvious point that no single methodology is applicable to all problems. It does seem useful (in some cases) to carefully test treatments that represent a stronger intervention than could be replicated nationally, in order to establish an upper bound on the effects that can reasonably be expected from a particular type of policy. I have no way of knowing whether that was part of the motivation for MRFIT.

Harris levels a similar criticism at the Health Insurance Study. As he indicates, that experiment was designed primarily to estimate the effects of alternative levels and forms of cost sharing on the demand for medical care. Thus, he argues, it excludes a variety of institutional and supply-side responses that might have an important effect on outcomes in a

national program, and therefore will not be able to directly predict those outcomes. I certainly would not quarrel with that characterization of the experiment, nor would its other designers. But I don't particularly regard that as a serious criticism either of the methodology or of the experiment itself. No single project can address all aspects of a complex, multi–billion dollar national program, and the Health Insurance Experiment is no exception. It was never intended to directly predict the utilization outcomes under any particular national health-insurance plan. Rather, it was intended to fill a gaping hole in our knowledge about underlying consumer behavior in the health sector. It was recognized from the beginning that demand-side information would have to be combined with whatever analyses are possible of institutional and supply-side response to derive national estimates of costs and utilization; but it is equally true that no reasonable estimates of national outcomes can be produced without the demand-side information that will be produced by the experiment. In short, this criticism is a perfect illustration of my fundamental disagreement with Harris over whether experiments should attempt holistic replication of complex policies or should simply attempt to generate reliable information on one or more—presumably important—pieces of the policy design problem.

Harris's discussion of macroexperiments is much more sanguine, although he does acknowledge some of the problems posed by this type of experiment. He notes, for example, that since the basic unit of observation in a macroexperiment is an entire group or community, the feasible sample size (i.e., number of communities) and representativeness may be severely limited, and that confounding of site and treatment effects may be a serious problem. It may also be difficult to randomize groups or communities to treatment and control status because of administrative or political considerations. Finally, he concedes that attempts to measure outcomes with longitudinal, individual-level survey data will be subject to many of the problems encountered in microexperiments and suggests that repeated cross-sectional surveys be conducted instead. His discussion of these issues seriously underestimates their likely severity, however, and is overly sanguine about their proposed remedies. It also omits some of the more serious difficulties of mounting rigorous macroexperiments.

Many of the problems of macroexperiments flow from the sheer size of the natural observational units. Where the unit of observation is an entire market, for example, these projects can be extremely expensive. The original planning budgets for HUD's housing-allowance supply experiment (two housing markets) and DOL's Employment Opportunity Pilot Projects (fifteen labor markets) were each on the order of $400 million. That is considerably more than the budgets of all the income-maintenance experiments, the health-insurance experiment, and the

housing-allowance demand experiment (all microexperiments) combined. The high cost of "saturating" an entire market leads, of course, to severe limits on the number of observations. It also tends to favor the selection of small markets and, for many purposes, virtually precludes selecting cities like New York, Los Angeles, or Chicago, thereby jeopardizing the representativeness of the sample.

The constraints on budget and sample size can be severe even with units of observation much smaller than an entire market. The Labor Department recently hired two contractors to prepare alternative designs for a set of demonstrations of employment and training services for workers disemployed in plant closings. One contractor estimated that the optimal number of plants required to disentangle plant effects from individual responses was 133; the other contractor (using different assumptions) arrived at an optimal sample size of about 1,000 plants. The DOL budget for the project was $50 million—about the cost of a "typical" microexperiment. That budget would have supported a sample of at most 50 plants, even if the sample were heavily skewed toward atypically small plants.

The sample-size constraint not only affects the statistical precision of the results, it also severely restricts the number of treatment options that can be tested. In the Employment Opportunity Pilot Projects, for example, an initial list of seven "planned variations" of the basic program was ultimately reduced to two, after long and painful deliberation, on the grounds that more variations within a fifteen-site project would jeopardize the chance of learning anything reliable about either the variations or the basic program itself. Even in the plant-closing demonstrations, with a potential sample of as many as fifty plants, both design contractors agreed that it would be risky to try more than five or six different treatments unless randomization within plants (i.e., embedding microexperiments within the macroexperiment) was allowed. In both of these cases, there were literally dozens or even hundreds of treatment levels and combinations that were of policy interest and very little ability to use the treatments actually tested to interpolate or extrapolate to options not tested.

These projects illustrate vividly the problem of black box experiments mentioned earlier. In projects like these, where only a small fraction of a large number of potential policy options can be implemented, the experimenter is in the almost impossible position of trying to predict which policy options will be relevant as much as ten years in the future, when the project has been completed and the data analyzed. In the light of the recent dramatic policy shifts at the federal level, this task seems almost hopeless. The Employment Opportunity Pilot Projects, for example, were focused heavily on public-service employment when they were initiated in 1979; in March 1981, President Reagan terminated all federal

support for public-service employment. By way of contrast, the treatments in the Health Insurance Experiment span the entire range of policy options in one important dimension of health-financing policy.

Perhaps the most serious problem arising out of the scale of macroexperiments is the difficulty of control and administration. It is not only that the number and size of sites required for valid inference presents a serious span-of-control problem—although that is certainly the case. The scale of these projects will often require that they be implemented by the regular-program bureaucracy. The experimenters' objectives will conflict in important ways with the objectives of regular-program operators, and it will be exceedingly difficult to ensure that even those few treatments selected for testing are actually implemented as intended. The monitors of DOL's Employment Opportunity Pilot Projects faced a steady stream of resistance and requests for exceptions to federal guidelines from the CETA prime sponsors running the project. Often program operators simply ignored the guidelines when they departed from normal practice or were in conflict with the operators' concept of what was best for the client, or for their own agency. One of the planned variations in the Employment Opportunity Pilot Projects was a set of employment subsidies designed to encourage placement of AFDC recipients in private-sector jobs. It was only after this subexperiment was well underway that the DOL monitors discovered that the only clients being referred to the project from the welfare agency were the rejects and failures from WIN's own placement activities. Needless to say, this had a major impact on the project's placement rate—the principal outcome measure—although it did wonders for WIN's placement rate.

The scale and visibility of macroexperiments also makes them extremely vulnerable to a variety of political pressures. The Minnesota Work Equity Project, for example, became embroiled in political controversy that delayed its implementation for nearly a year and had serious adverse effects on its ultimate design and implementation. The Employment Opportunity Pilot Projects, with an annual budget of $100 to $200 million, was an obvious target for federal budget cutters throughout its brief life. In 1980, the project was seriously scaled back in midcourse as part of President Carter's budget-balancing effort, and in 1981 it was prematurely terminated by the new administration. In contrast, the Health Insurance Study, with an annual budget less than one-tenth as large, escaped the budget cutters' ax on both occasions.

A final limitation of macroexperiments, not discussed by Harris, is closely analogous to the selectivity problem posed by voluntary participation in macroexperiments. Participation in macroexperiments is, after all, also voluntary—both at the site and individual levels. The experience with site selection for experiments and demonstrations is no more encouraging than the individual take-up rates in microexperiments. In the

Food Stamp Workfare Demonstrations, for example, a national solicitation netted a total of seven volunteer sites—six rural counties and one urban county where the food stamp caseload is allegedly heavily composed of "beach bums." In both the Health Insurance Study and the Employment Opportunities Pilot Projects, the experimenters first selected sites and then approached the local authorities, attempting to elicit their approval and/or cooperation. In both cases, exactly one-seventh of the sites approached either refused or failed to cooperate to such a degree that a program was never initiated.

The issue of individual participation and selectivity bias is somewhat more subtle in macroexperiments than in microexperiments, but no less real or important. In any intervention that relies for its effect on any positive action on the part of individuals, the extent of individual participation—whether it be enrollment in a program, application for benefits, or simply response to a mass media educational campaign—will be heavily dependent on the level of program outreach. The level and effectiveness of outreach efforts are exceedingly difficult to control in most cases, and the resulting participation rates can vary widely; I have seen participation rates anywhere from 1 or 2 percent to 40 or 50 percent in response to what appear to be comparable outreach efforts. The individuals who respond to outreach directed toward the general population are, of course, just as self-selected as the randomly selected individuals who agree to participate in a microexperiment. Indeed, since participation rates are likely to be much lower in a macroexperiment, the potential for selectivity bias seems more serious. If one had confidence that the experimental outreach and participation would be replicated in a national program (or in another site), this potential bias would not be a problem, since the experimental outcomes would then be unbiased predictors of national-program outcomes. But that seems a heroic assumption, given the idiosyncratic nature of local outreach activities and the extreme variation in resulting participation rates.

While I do not believe that macroexperiments avoid the selectivity-bias problem, I do feel that one of the advantages of macroexperiments is their potential for measuring, if only crudely, participation rates. For the reasons just discussed, participation rates in a macroexperiment are likely to be an imprecise predictor of national rates, but participation can't be predicted at all from a microexperiment because the outreach method is highly artificial. And participation rates are a very important determinant of program cost and/or effectiveness.

Harris suggests several methodological approaches to mitigate the shortcomings of macroexperiments. I agree that steps could be taken to improve the methodological rigor of such projects. I am more skeptical than Harris, however, as to the practicality and likely effectiveness of some of his suggestions.

I heartily endorse, for example, the suggestion of random selection of treatment and comparison sites from matched pairs. I am much more dubious about the possibility of "crossover and fold-back designs" in which the timing and sequence of program start-up and termination are "random and unknown to the experimental units." In most cases projects like these require extensive prior negotiation and planning with local officials or agencies; it would often be virtually impossible—and possibly unethical—to keep such crucial information from the local personnel.

On the other hand, Harris's suggestion of mixed macro and micro designs is quite appealing. In fact, the design of the Employment Opportunity Pilot Projects included two microexperiments embedded within the overall macroexperiment. Unfortunately, the results of that effort were not entirely encouraging, largely because the microexperiments, like the macroexperiment itself, were administered by the regular program operators; the challenge of implementing random assignment and multiple treatments proved to be a difficult one for the program operators. I have already described the problems encountered in sample referral from WIN to the employment-subsidy experiment; the other experiment-within-the-experiment, involving alternative job-search assistance techniques, had such difficulty establishing an effective outreach effort that the results were virtually useless, and ultimately the experiment was abandoned.

The suggestion that the results of many independent macroexperiments could be combined is less appealing, even in principle. My office has just completed a survey of about a dozen experiments and demonstrations in job-search assistance, all of which were modeled on a single, apparently successful, project. While we did not attempt any rigorous pooling of data or results, it quickly became clear that the diversity of treatment design, data collection, outcome definition, and sample selection in these projects almost defied description, let alone formal modeling. I am doubtful that the task would be any easier in most other cases.

Finally, Harris suggests that many of the problems of data collection could be avoided by using repeated cross-sectional surveys, rather than longitudinal surveys. While there is some validity to this suggestion, the precision of the estimates of treatment effects could suffer substantially because of individual variation. Moreover, contrary to his assertion, repeated cross sections might be much more expensive in many cases than longitudinal surveys. If the population of interest is a subset of the general population (e.g., poor people, sick people, or program eligibles) a large number of screening interviews may be required to identify each useful observation. In a longitudinal survey, this screening operation need only be performed once; in repeated cross sections it would have to be done for each successive wave. In the Employment Opportunity Pilot Projects, for example, approximately fifteen screening interviews with a

random sample of the general population were required for each program eligible identified.

Taking all of the strengths and weaknesses of both experimental modes into account, I would propose the following general criteria for deciding which experimental method is appropriate for a particular policy issue.

1. *Policy interest.* If the objective is to measure the overall efficacy of a single program or small number of programs, macroexperimentation may be more appropriate; if the objective is to estimate behavioral responses to a wide range of program variants, microexperimentation is indicated. Policy interest in estimation of participation rates also favors macroexperimentation.
2. *Nature of the treatment.* The nature of the treatment will occasionally dictate one mode of experimentation. For example, educational campaigns that rely on mass media could not be implemented in a microexperiment. On the other hand, treatments that require complicated explanations or interactions with participants—such as "buying out" an existing health-insurance plan—may be better implemented in a microexperiment.
3. *Nature of the response.* Macroexperimentation may be indicated if interactions among individuals in the group or community are thought to have an important effect on the response to the treatment. Purely individualistic responses can be measured in either mode of experiment.
4. *Administrative considerations.* The scale and objectives of macroexperiments will usually dictate that they be administered through existing institutions and organizations. Careful thought must be given to whether that is possible in a experimental context. Span-of-control problems, competing institutional objectives, ingrained organizational behavior, and garden-variety start-up problems may seriously compromise implementation of the treatment in a short-duration experiment. On the other hand, microexperiments will be even more difficult to run through existing institutions because of the multiplicity and complexity of treatments; they will usually require a special administrative structure under the direct control of the experimenters. The degree to which this arrangement realistically replicates the administrative structure of a permanent program, and how critical the difference is to the outcomes of interest, must be carefully assessed.
5. *Statistical considerations.* The estimates of treatment effect are likely to be more precise and unbiased in a microexperiment, because of the problems of cost, sample size, selectivity bias, administrative control, and lack of a true control group in macroexperiments. These potential disadvantages must be analyzed and weighed against whatever other factors favor macroexperimentation.

References

Arrow, Kenneth J. 1975. *Two notes on inferring long-run behavior from social experiments.* Rand Report P–5546. Santa Monica: Rand Corporation.

Apostolides, Aristide, and Maureen Henderson. 1977. Evaluation of cancer screening programs: Parallels with clinical trials. *Cancer* 39: 1179–85.

Brook, Robert H., John E. Ware, Jr., Allyson Davies-Avery, et al. 1979. *Conceptualization and measurement of health for adults in the health insurance study: Vol. 7, overview.* Rand Report R–1987/8–HEW. Santa Monica: Rand Corporation.

Campbell, Donald T., and Julian C. Stanley. 1966. *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand McNally.

Cornfield, Jerome. 1978. Randomization by group: A formal analysis. *American Journal of Epidemiology* 108: 100–102.

Davis, C. E., and R. J. Havlik. 1977. Clinical trials of lipid lowering and coronary artery disease prevention. In *Hyperlipidemia: Diagnosis and therapy*, ed., B. M. Rifkind and R. I. Levy. New York: Grune and Stratton.

Duan, Naihua, Willard G. Manning, Carl N. Morris, and Joseph Newhouse. 1983. A comparison of alternative models of the demand for medical care. *Journal of Business and Economic Statistics* 1: 115–26.

DuMouchel, William H., and Jeffrey E. Harris. 1983. Bayes and empirical Bayes methods for combining the results of cancer studies in humans and other species (with discussion). *Journal of the American Statistical Association* 78: 293–315.

Farquhar, John W. 1978. The community-based model of life-style intervention trials. *American Journal of Epidemiology* 108: 103–11.

Farquhar, John W., Nathan Maccoby, Peter D. Wood, et al. 1977. Community education for cardiovascular health. *Lancet* 1: 1192–95.

Feldstein, Martin S. 1977. Quality change and the demand for hospital care. *Econometrica* 45: 1681–702.

Friedman, Gary D., Diana B. Petitti, Richard D. Bawol, and A. B. Siegelaub. 1981. Mortality in cigarette smokers and quitters. *New England Journal of Medicine* 304: 1407–10.

Gillum, Richard F., Paul T. Williams, and Edward Sondik. 1980. Some considerations for the planning of total-community prevention trials: When is sample size adequate? *Journal of Community Health* 5: 270–78.

Harris, Jeffrey E. 1982. Prenatal medical care and infant mortality. In *Economic Aspects of Health*, ed. V. Fuchs. Chicago: University of Chicago Press.

———. 1980. Commentary. *National Health Insurance: What now? What later? What never?*, ed. M. Pauly. Washington, D.C.: American Enterprise Institute.

———. 1979. The aggregate coinsurance rate and the supply of innovations in the hospital sector. Department of Economics working paper, Massachusetts Institute of Technology.

Hulley, Stephen B., and Stephen F. Fortman. 1980. Clinical trials of changing behavior to prevent cardiovascular disease. In *Perspectives in behavioral medicine*, ed. S. M. Weiss. New York: Academic Press.

Hypertension Detection and Follow-up Program Cooperative Group. 1979a. Five-year findings of the hypertension detection follow-up program: I. Reduction in mortality of persons with high blood pressure, including mild hypertension. *Journal of the American Medical Association* 242: 2562–71.

Hypertension Detection and Follow-up Program Cooperative Group. 1979b. Five-year findings of the hypertension detection follow-up program: II. Mortality by race, sex, and age. *Journal of the American Medical Association* 242: 2572–77.

Kasl, Stanislav V. 1980. Cardiovascular risk reduction in a community setting: Some comments. *Journal of Consulting and Clinical Psychology* 48: 143–49.

———. 1978. A social-psychological perspective on successful community control of high blood pressure: A review. *Journal of Behavioral Medicine* 1: 347–81.

Keeler, Emmett B., John E. Rolph, Naihua Duan, et al. 1982. *The demand for episodes of medical treatment: Interim results from the Health Insurance Experiment.* Rand Report R–2829–HHS. Santa Monica: Rand Corporation.

Kuller, Lewis, James Neaton, Arlene Caggiula, and Lorita Falvo-Gerard. 1980. Primary prevention of heart attacks: The multiple risk factor intervention trial. *American Journal of Epidemiology* 112: 185–99.

Leventhal, Howard, Martin A. Safer, Paul D. Cleary and Mary Gutman. 1980. Cardiovascular risk modification by community-based programs for life-style change: Comments on the Stanford study. *Journal of Consulting and Clinical Psychology* 48: 150–58.

Maccoby, Nathan, John W. Farquhar, Peter D. Wood, and Janet Alexander. 1977. Reducing the risk of cardiovascular disease: Effects of a community-based campaign on knowledge and behavior. *Journal of Community Health* 3: 100–114.

Manning, Willard G., Jr., Carl N. Morris, Joseph P. Newhouse, et al. 1981. A two part model of the demand for medical care: Preliminary results from the Health Insurance Study. In *Health, economics, and*

health economics, ed. J. van der Gaag and M. Perlman. Amsterdam: North Holland.

Manning, Willard G., Jr., Joseph P. Newhouse, and John E. Ware, Jr. 1982. The status of health in demand estimation: Beyond excellent, good, fair, and poor. In *Economic Aspects of Health*, ed. V. Fuchs. Chicago: University of Chicago Press.

Meyer, Anthony J., Joyce D. Nash, Alfred L. McAlister, Nathan Maccoby, and John W. Farquhar. 1980. Skills training in a cardiovascular health education campaign. *Journal of Consulting and Clinical Psychology* 48: 129–42.

Morris, Carl. 1979. A finite selection model for experimental design of the Health Insurance Study. *Journal of Econometrics* 11: 43–61.

Morris, Carl N., Joseph P. Newhouse, and Rae W. Archibald. 1980. *On the theory and practice of obtaining unbiased and efficient samples in social surveys*. Rand Report R–2173–HEW. Santa Monica: Rand Corporation.

Mosteller, Fred, and Gail Mosteller. 1979. New statistical methods in public policy: Part I, experimentation. *Journal of Contemporary Business* 8: 79–92.

Multiple Risk Factor Intervention Trial Group 1982. Multiple risk factor intervention trial: Risk factor changes and mortality results. *Journal of the American Medical Association* 248: 1465–77.

———. 1977. Statistical design considerations in the NHLBI Multiple Risk Factor Intervention Trial. *Journal of Chronic Diseases* 30: 261–75.

———. 1976a. The Multiple Risk Factor Intervention Trial (MRFIT). *Journal of the American Medical Association* 235: 825–27.

———. 1976b. The Multiple Risk Factor Intervention Trial. *Annals of the New York Academy of Medicine* 304: 293–308.

Newhouse, Joseph P. 1978. *The erosion of the medical marketplace*. Rand Report R–2141. Santa Monica: Rand Corporation.

———. 1974. A design for a health insurance experiment. *Inquiry* 2: 5–27.

Newhouse, Joseph P., Willard G. Manning, Carl N. Morris, et al. 1982. *Some interim results from a controlled trial of cost sharing in health insurance*. Rand Report R–2847–HHS. Santa Monica: Rand Corporation.

Newhouse, Joseph P., Kent H. Marquis, Carl N. Morris, Charles E. Phelps, and William H. Rogers. 1979. Measurement issues in the second generation of social experiments: The Health Insurance Study. *Journal of Econometrics* 11: 117–29.

Ockene, Judith K., Norman Hymowitz, Mary Sexton, and Steven K. Broste. 1982. Comparison of patterns of smoking behavior change

among smokers in the Multiple Risk Factor Intervention Trial. *Preventive Medicine* 11: 621–38.

Puska, P., J. Tuomilehto, A. Nissinen, et al. 1978. Changing the cardiovascular risk in an entire community: The North Karelia project. Paper presented at the International Symposium on Primary Prevention in Early Childhood of Atherosclerotic and Hypertensive Diseases, Chicago, Ill.

Rivlin, Alice. 1974. Allocating resources for policy research: How can experiments be more useful? *American Economic Review Papers and Proceedings* 64: 346–54.

Rose, Geoffrey, and R. J. S. Hamilton. 1978. A randomised controlled trial of the effect on middle-aged men of advice to stop smoking. *Journal of Epidemiology and Community Health* 32: 275–81.

Rose, Geoffrey, R. F. Heller, Hugh T. Pedoe, and D. G. S. Christie. 1980. Heart disease prevention project: A randomized controlled trial in industry. *British Medical Journal* 280: 747–51.

Schoenberger, James A. 1981. The Multiple Risk Factor Intervention Trial. Presentation at American Heart Association meetings, Washington, D.C.

Sherwin, Roger. 1978. Controlled trials of the diet-heart hypothesis: Some comments on the experimental unit. *American Journal of Epidemiology* 108: 92–99.

Sherwin, Roger, Mary Sexton, and Patricia Dischinger. 1979. The Multiple Risk Factor Intervention Trial of the primary prevention of coronary heart disease: Risk factor changes after two years. Paper presented at the Seventh Asian Pacific Congress of Cardiology, Bangkok.

Stern, Michael P., John W. Farquhar, Nathan Maccoby, and Susan H. Russell. 1976. Results of a two-year health education campaign on dietary behavior: The Stanford three community study. *Circulation* 54: 826–33.

Strolley, Paul D. 1980. Epidemiologic studies of coronary heart disease: Two approaches. *American Journal of Epidemiology* 112: 217–24.

Syme, S. Leonard. 1978. Life style intervention in clinic-based trials. *American Journal of Epidemiology* 108: 87–91.

Truett, J., J. Cornfield, and W. Kannel. 1967. Multivariate analysis of the risk of coronary heart disease in Framingham. *Journal of Chronic Diseases* 20: 511–24.

Ware, John E., Jr., Robert H. Brook, Allyson Davies-Avery, et al. 1980. *Conceptualization and measurement of health for adults in the health insurance study: Vol. 1, model of health and methodology*. Rand Report R–1987/1–HEW. Santa Monica: Rand Corporation.

WHO European Collaborative Group. 1974. An international controlled trial in the multifactorial prevention of coronary heart disease. *International Journal of Epidemiology* 3: 219–24.

Williams, Paul T., Stephen P. Fortmann, John W. Farquhar, Ann Varady, and Susan Mellan. 1981. A comparison of statistical methods for evaluating risk factor changes in community-based studies: An example from the Stanford three-community study. *Journal of Chronic Diseases* 34: 565–71.